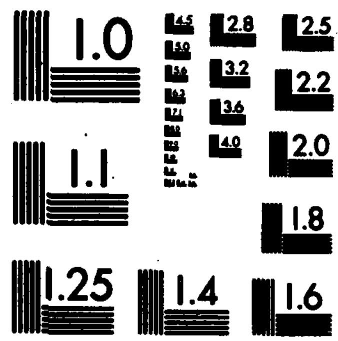
AD-A147 872 CONDITIONAL REGRESSION MODELS FOR TRANSIENT STATE SURVIVAL ANALYSIS..(U) NORTH CAROLINA UNIV AT CHAPEL HILL INST OF STATISTICS R D ABBOTT ET AL. AUG 84 1/1

UNCLASSIFIED MIMEO SER-1555 AFOSR-TR-84-0951 F/G 12/1 NL

END

FILMED

DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD-A147 872

CONDITIONAL REGRESSION MODELS FOR

TRANSIENT STATE SURVIVAL ANALYSIS

Robert D. Abbott
Field Studies Branch
National Heart, Lung and Blood Institute
National Institutes of Health

Raymond J. Carroll
Department of Statistics
University of North Carolina
Chapel Hill, NC 27514

Accession For	
NTIS GRA&I	X
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

Professor Carroll's work was supported by the Air Force Office of Scientific Research AFOSR F 49620-82-C-0009.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): Mimeo Series #1555	5. MONITORING ORGANIZATION REPORT NUMBER(S): AFOSR-TR- 84_0951
6a. NAME OF PERFORMING ORGANIZATION: University of North Carolina; 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION: Air Force Office of Scientific Research
6c. ADDRESS (City, State and ZIP Code): Department of Statistics, Chapel Hill NC 27514	7b. ADDRESS (City, State and ZIP Code): Directorate of Mathematical & Information Sciences, Bolling AFB DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: AFOSR; 8b. OFFICE SYMBOL (If applicable): NM	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: F49620-82-C-0009
8c. ADDRESS (City, State and ZIP Code): Bolling AFB DC 20332-6448	10. SOURCE OF FUNDING NOS: PROGRAM ELEMENT NO 61102F; PROJECT NO 2304; TASK NO A5; WORK UNIT NO

11. TITLE (Include Security Classification): CONDITIONAL REGRESSION MODELS FOR TRANSIENT STATE SURVIVAL ANALYSIS

12. PERSONAL AUTHOR(S): Robert D. Abbott* and Raymond J. Carroll

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
Technical	FROM ______ TO ______	AUG 84	12

16. SUPPLEMENTARY NOTATION: *Field Studies Branch, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda MD.

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB. GR.	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Survival models are important tools for the analysis of data when a disease event occurs with time and subjects are lost to follow-up. Many models, however, can also be adapted for use when an event is characterized by transitions through intermediate states of disease with increasing severity. In this presentation, such an adaptation will be demonstrated for a class of conditional regression models for the analysis of transient state events occurring among grouped event times. The type of conditioning that will be described is useful in providing comparisons of specific disease states and an assessment of transition dependent risk factor effects. An example will be given based on the Framingham Heart Study.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: UNCLASSIFIED/UNLIMITED X SAME AS RPT. ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION: UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL: MAJ Brian W. Woodruff	22b. TELEPHONE NUMBER (Include Area Code): (202) 767-5027; 22c. OFFICE SYMBOL: NM

DD FORM 1473, 83 APR

EDITION OF 1 JAN 73 IS OBSOLETE.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

ABSTRACT

Survival models are important tools for the analysis of data when a disease event occurs with time and subjects are lost to follow-up. Many models, however, can also be adapted for use when an event is characterized by transitions through intermediate states of disease with increasing severity. In this presentation, such an adaptation will be demonstrated for a class of conditional regression models for the analysis of transient state events occurring among grouped event times. The type of conditioning that will be described is useful in providing comparisons of specific disease states and an assessment of transition dependent risk factor effects. An example will be given based on the Framingham Heart Study.

Much has been written about the analysis of survival data when subjects are followed over a period of time for the development of a disease event. Important models have emerged that have enabled the examination of risk factor relationships to disease while also considering that subjects are often lost to follow-up (1-4).

The models, however, are usually limited to the analysis of data when subjects take single jumps from an event free state to a state of event development. Despite this limitation, it is easy to envision that the single jumps that are modeled can often be characterized by an underlying disease process. It is more likely that the single disease transition can be better described by intermediate jumps. Furthermore, the relationship of risk factors to disease may depend on the type of transition being made. Some risk factors may have a greater influence on promoting one transition than on others.

For example, cholesterol may be more strongly related to the first appearance of coronary heart disease than it is to secondary events (5,6). As a result, the relationship of cholesterol to different transitions in coronary heart disease development (ranging from healthy, to intermediate states of disease, to death) may not be the same. On the otherhand, due to myocardial irritability, cigarette smoking may have a more profound effect on the abrupt jump from clinically healthy to sudden unexpected death (7,8). Clearly, if risk factor effects emerge differently in the genesis of disease, improved modeling of these different effects to better explain disease development is necessary.

One approach would be to model the probability of making transitions to various disease states given an initial state and one or more risk factors. Care should be taken, however, to define these probabilities appropriately to take into account the transitory effects of the risk factors.

This report will present some ideas of how this can be accomplished. The ideas will be based on modeling multiple transitions and risk factor effects on the genesis of disease. A general class of conditional regression models for transient state survival analysis will be presented based on grouped event times and a discrete and ordered process of disease development. Although other models exist for the analysis of transient disease states (9-11), the ideas considered here are different since they provide a unique kind of conditioning other than those that usually involve risk factors or the quality of pre-existing disease. Such conditioning will be shown to permit a useful comparison of specific disease states and a description of transition dependent risk factor effects. An example based on data from the Framingham Heart Study will be given that is particularly suited for this kind of analysis.

METHODS

Consider a set of risk factors, $x_1, x_2, \ldots x_r$, that are measured on a sample of subjects at some baseline. Assume that for each subject, time intervals ($t = 1,2,\ldots\tau$) are observed when a disease endpoint occurs or when a subject is lost to follow-up. Let $Y_t = 1$ denote the occurrence of an event for a subject during an interval of time indexed by t and $Y_t = 0$ otherwise.

A general expression for a class of survival models for the analysis of this kind of data when event times are grouped is given as follows:

$$\text{link}(p_T) = \alpha_T + \Sigma\beta_k x_k. \qquad (1)$$

In the above model,

$$p_T = \Pr(Y_T = 1 \mid Y_t = 0,\ t < T). \qquad (2)$$

Here, p_T is the probability of developing the disease during the interval of time indexed by T among subjects who are healthy prior to that interval.

The link in model 1 represents a transformation of p_T with tractable features, some of which will be described later. In logistic regression, the link is the logit transform and corresponds to the model of Wu and Ware (3). When the proportional hazards model of Cox holds, the complementary log(-log) transform is the link. This latter link is attributed to Prentice and Gloecker (2). Other links are described by McCullagh (12). For a further discussion, see Abbott (13).

The coefficient α_T in model 1 is an intercept unique to the interval T. The coefficient β_k is, of course, a measure of association between the kth risk factor and the probability of developing disease.

Suppose now that the above survival experiences can be characterized by a subjects movement into ordered and discrete states of disease severity. Assume further, that transitions to healthier states are not possible, as is usually the case in chronic disease epidemiology. Suppose that the disease states are indexed by $i = 1,2...k$ where the index i decreases with increasing severity. Let $Y_{i|t} = 1$ if the ith event is the worst event occurring in the time interval t and $Y_{i|t} = 0$ otherwise. Model 1 is appropriate for analyzing such survival experiences, but it needs to be adapted to enable modeling disease severity.

To develop such an adaptation, it is natural to define the following probability for $J < I$.

$$\Pr(Y_{j|T} = 1 \text{ for some } j \le J \mid Y_{I|T-1} = 1) \tag{3}$$

Expression 3 represents the probability that a subject falls into a disease state $J < I$ or worse during interval T among those who were in state I during interval T-1. Such a probability is useful to consider. Because disease states are pooled, however, the chance of moving into a specific disease state cannot be examined directly. By taking the appropriate difference, however, one could define for $J < I$

$$Pr(Y_{J|T} = 1 \mid Y_{I|T-1} = 1). \tag{4}$$

Expression 4 represents the probability that a subject in disease state I during interval T-1 moves exactly into disease state J during interval T. Suppose, however, that J represents some intermediate state of disease between healthy and death. Then the resulting probability intrinsically compares subjects who remain healthy or die as a single group with those who move into the intermediate disease state. This probability provides a peculiar comparison. It is perhaps more sensible if it were conditioned on events worse than J not occurring.

As a result, consider the following for $J < I$.

$$p_{IJ|T} = Pr(Y_{J|T} = 1 \mid Y_{I|T-1} = 1 \text{ and } Y_{j|T} = 0,\ j < J) \tag{5}$$

Here, $p_{IJ|T}$ is the probability that a subject falls in the disease state $J < I$ during interval T among subjects who 1) were in the Ith state at T-1 and 2) who experienced disease no worse than J during the interval T. Such a probability now describes the chance of moving from an initial state I to a specific state of interest, J, without being obscured with movements into states more severe than J.

An adaptation to model 1 that considers the conditional transition probabilities 5 as a function of risk factors is given as follows:

$$\text{link}(p_{IJ|T}) = \mu_T + \delta_{IJ} + \Sigma\beta_{IJ|k}x_k. \tag{6}$$

Although $p_{IJ|T}$ can be expressed in terms of the probabilities given by 3 and 4, it is easier to write $p_{IJ|T}$ directly in the form given by model 6 to simplify interpretation of the risk factor coefficients. Further discussion on this will appear later.

Model 6 is called a conditional regression model because it has the added feature of being conditioned on an event worse than J not occurring. This latter condition insures that the effects of the risk factors on the transition from state I to J are not confused with movements into states more severe than J.

The parameters δ_{IJ} and $\beta_{IJ|k}$ are now associated with the transition from the Ith to the Jth state. As with model 1, μ_T is an intercept unique to the time interval T. Clearly, the relationship of the parameters to a transition has special meaning depending on the link being used. For example, when the logistic transform is the link, the conditioned model implies that the odds, $p_{IJ|T}/(1-p_{IJ|T})$, is proportional to $\exp(\mu_T)$ where the proportionality constant is $\exp(\delta_{IJ}+\Sigma\beta_{IJ|k}x_k)$.

As in any analysis of variance type situation, the conditional model can be generalized by replacing $\mu_T+\delta_{IJ}$ with $\alpha_{IJ|T}$. Such a generalization allows inclusion of interaction effects between the transition types and time. As is often the case, however, the resulting number of parameters can be quite large and beyond the limits of the data.

Notice further that the conditional model assumes that the transitory effects of the risk factors do not depend of T. Nevertheless, the model can also be generalized to include time dependent covariates at the cost of adding extra parameters. Wu and Ware (3) provide ideas of how this can be accomplished.

Estimation of the coefficients in the conditional model is accomplished numerically by maximizing the likelihood of the data. The likelihood is constructed by the first conditioning on the Ith state of disease a specific subject experiences during the interval T-1. For the interval of time indexed by T, the contribution by the subject to the likelihood is then;

$$p_{IJ|T}^{\theta} \prod_{j < J} (1-p_{Ij|T}). \tag{7}$$

where $\theta = 1$ if a subject develops the Jth level of disease severity where $J < I$ and $\theta = 0$ if the disease state remains unchanged; i.e., $J = I$. The entire contribution to the likelihood by an individual is then the product of all relevant terms 7 over all time intervals for which loss to follow-up has not occurred or a transition to another disease state is still possible. The complete likelihood is then the product of all individual contributions. As expected, in the single transition setting, the resulting likelihood reduces to expressions given elsewhere (2-3,13).

A description of some useful ideas for maximizing 7 when the complementary log(-log) transform is the link in the conditional model is given by Prentice and Gloeckler (2). Maximizing the likelihood using the logit transform is accomplished by following the ideas of Wu and Ware (3). Wu and Ware also show how to maximize their likelihood using a packaged logistic regression routine (14). With little effort, these guidelines can be easily extended to estimating the parameters in the conditional model 6.

Depending on the link used, interpreting the coefficients in the conditional model is simple. For example, suppose that two individuals are characterized by risk factors x_k and x_k', respectively, $k = 1,2,\ldots r$. When the rate of disease is small and either the complementary log(-log) or logistic transforms are the links being used, the relative risk of moving from state I to J between these two subjects, conditioned on not experiencing a transition to a state worse than J, is approximately $\exp[\Sigma\beta_{IJ|k}(x_k-x_k')]$. See Abbott (13) for a further discussion.

AN EXAMPLE

To illustrate use of the conditional model, data collected from the Framingham Heart Study (15) is examined. The information is from 476 males who were aged 40-49 years and free of coronary heart disease when first examined beginning in 1948. The risk factors used were measured at the time of the initial exam and include systolic blood pressure, a body mass index measured as weight/height^2, total cholesterol, and the use of cigarettes. Age was not considered as controlling information in the example since its range was restricted and proved to be insignificant in the subsequent analyses.

The endpoint of interest is coronary heart disease. The follow-up for this event after the first recruitment date consists of 13 biennial exams. During this period, there were 147 cases of coronary heart disease, among which, 62 deaths were observed.

In this example, the adaptation that models discrete and ordered states of disease severity is useful since three states of disease are considered which possess these features. The states include clinically healthy ($I=3$), alive with coronary heart disease ($I=2$), and death from coronary heart disease ($I=1$). Also, the adaptation that considers grouped event times is an appropriate model simply by the nature of one of the states considered. For example, in the data, alive with coronary heart disease is not always a clearly defined event in terms of when it actually occurs. Many times, coronary heart disease is manifested as angina pectoris or an unrecognized myocardial infarction, and a diagnosis is often made only during a routine scheduled physical exam. As a result, the diagnosis alive with coronary heart disease is assigned in the Framingham Study to the subjects exam at which it was first detected. Since the exact time disease developes is not always known, neither will the time from its first appearance to death from coronary heart disease be known. Thus, it is natural to group all event times to the interval of time where it was most likely to occur.

For this particular example, the logistic link will be used in the conditional model. Its relationship with the complementary log(-log) link will be described later.

The resulting risk factor coefficients from the conditional model using the logistic link are given in table 1. Notice that body mass significantly influences the transition from healthy to alive with coronary heart disease ($p = 0.009$). Systolic blood pressure also seems to influence this transition ($p = 0.058$) with a similar affect on the transition from healthy to death from coronary heart disease ($p = 0.047$). Cigarette use is also an important contributor to this latter transition ($p = 0.009$) with a less marked affect existing for cholesterol ($p = 0.097$). None of the risk factors significantly promotes the transition to death once a subject has already developed coronary heart disease, a possible consequence of small numbers.

DISCUSSION

The purpose of this paper is to indicate the ability of regression models to consider facets of survival analysis other than single transitions between two states. For the particular data presented, a conditional model is sensible to consider, because the three types of transitions provide different descriptions of risk factor effects on pathogenesis. It is apparent that modeling these transitions suggests the possibility that the risk factors contribute to disease states differently; a concept that is not new (5-8). Certainly the strength of some risk factor relationships with disease are stronger than others depending on the transitions being made. For example, body mass is strongly related to the transition from healthy to alive with coronary heart disease, while systolic blood pressure and cigarette smoking are related to the abrupt two year jump from clinically healthy to death. Much of this latter transition is associated wth sudden death from coronary heart disease.

Of course, it would have been natural to consider models that make use of such transition probabilities as those given by expression 3. The resulting analysis would produce findings that relate risk factor effects involving jumps to pooled states of disease. The interest in this presentation, however, is to determine if risk factors influence the transition between specific disease states. This is accomplished by conditioning on a state I at the beginning of a time interval and then estimating the chance of moving into a more severe state $J < I$ in the next interval. To insure that the effects of the risk factors on such a transition are not confused with movements into states more severe than J, an additional condition is made that events worse than J do not occur.

At this point, it may be apparent that simplifications of the conditional model 6 can be considered when analyzing the Framingham data. Clearly, the transition from alive with disease to death has offered little in explaining the relationship of risk factors to this particular transition, at least in terms of significance. It is not surprising that once disease develops, that past values of the risk factors do not discriminate well between survival and death. This finding may mean that when someone develops coronary heart disease, sufficient damage to the subjects health has occurred to result in a poor prognosis regardless of prior risk factor status. Although the Framingham data describing this transition is sparse, such a notion is useful to consider. Of course, one could always eliminate this transition from the analysis.

An alternative idea would be to reduce the conditional model by assuming that various transitions occur at similar rates. In the Framingham data, this would involve testing the hypothesis H_0: $\delta_{31} = \delta_{32} = \delta_{21}$. As expected, the result of such a test is sugnificant ($p < 0.001$). Clearly, this suggests that the rate of transition depends on the type of transition being made.

One can also examine if the coefficients for a specific risk factor are the same across transitions; i.e., test H_0: $\beta_{31|k} = \beta_{32|k} = \beta_{21|k}$. In the Framingham data, testing this hypothesis for systolic blood pressure, body mass, and cholesterol does not lead to a significant reduction in the likelihood of the data. In fact, the resulting pooled coefficients from the reduced model all become positive and significant ($p < 0.05$). Assuming that the cigarette coefficients are the same, however, has a significant affect on the likelihood ($p < 0.05$), and thus the coefficients should remain transition dependent. This latter finding is largely attributed to the distinct affect that smoking has on the jump from healthy to death.

Of course, the data analysis can be simplified further by using ordinary survival models. Such an analysis would require pooling adjacent disease states leading to different interpretations of risk factor relationships with disease. The resulting regression coefficients would then refer to the single state transition defined by this pooling. Indeed, the significance of some risk factor relationships can be determined by the transition being modeled. Such results are certainly useful to consider, but the models used to generate these findings are completely different from equation 6.

To further generalize the presented ideas, it is easy to envision competitors of the logistic link used in the example. The proportional hazards model for grouped event times is an important alternative (2). When disease occurs at slow rates, however, the results from such a model will differ only slighly from the logistic model used here (13). The proportional hazards model for grouped event times was actually applied to the Framingham example and the results were similar to those given by the logistic link.

Other link functions can be easily created. They should, however, usually be monotone, map the zero-one interval to the real number line, or enable easy interpretation of coefficients. If simultaneously modeling event times and disease transitions is also important, the link should possess the additional feature of being adaptable to these interests.

REFERENCES

1. Cox, D. Regression models and life tables (with discussion). J. Royal St. Soc., B 1972; 34: 187-220.

2. Prentice, R. and Gloeckler, L. Regression analysis of grouped survival data with application to breast cancer data. Biometrics 1978; 34: 57-67.

3. Wu, M. and Ware, J. On the use of repeated measurements in regression analysis with dichotomous responses. Biometrics 1979; 35: 513-21.

4. Bennett, S. Log-logistic regression models for survival data. Appl. Stat. 1983; 32: 165-71.

5. The Coronary Drug Project Research Group. Factors influencing long-term prognosis after recovery from myocardial infarctions - Three year findings of the coronary drug project. J. Chron. Dis. 1974; 27: 267-85.

6. Davis, C. and Havlik, R. Clinical trials of lipid lowering and coronary artery disease. In: Rifkind, B., Levy, R., eds. Hyperlipidemia: Diagnosis and therapy. New York, NY: Grune and Stratton, 1977: 79-92.

7. Kannel, W. Update on the role of cigarette smoking in coronary artery disease. Am. Heart J. 1981; 101: 319-28.

8. Schatzkin, A. Cupples, A., Heeren, T., et al. The epidemiology of sudden unexpected death: Risk factors for men and women in the Framingham Heart Study. Am. Heart J. 1984; 107: 1300-6.

9. Lagakos, S., Sommer, C. and Zelen, M. Semi-Markov models for partially censored data. Biometrika 1978; 65: 311-17.

10. Temkin, N. An analysis of transient states with application to tumor shrinkage. Biometrics 1978; 34: 571-80.

11. Voelkel, J. and Crowley, J. Nonparametric inference for a class of semi-Markov processes with censored observations. Ann. Stat. 1984; 12: 142-60.

12. McCullagh, P. Regression models for ordinal data (with discussion). J. Royal Stat. Soc., B 1980; 42: 109-42.

13. Abbott, R. Logistic regression in survival analysis. Am. J. Epid. (to appear).

14. Harrel, F. The logist procedure. In: Reinhardt, P., ed. SAS Supplemental Library User's Guide. Cary, NC: SAS Institute Inc., 1980: 83-89.

15. Dawber, T. The Framingham Study. Cambridge, MA: Harvard University Press, 1980.

TABLE 1

Risk factor coefficients in a conditional logistic regression model for a transient state survival analysis of coronary heart disease

Transition from healthy to death from coronary heart disease

Risk factor	Coefficient	Standard error	p-value
Systolic blood pressure	0.0153	0.0077	0.047
Body mass	0.0240	0.0485	0.621
Cholesterol	0.0049	0.0030	0.097
Smoking	1.2780	0.4879	0.009

Transition from healthy to alive with coronary heart disease

Risk factor	Coefficient	Standard error	p-value
Systolic blood pressure	0.0094	0.0050	0.058
Body mass	0.0798	0.0304	0.009
Cholesterol	0.0019	0.0023	0.394
Smoking	0.3110	0.2237	0.164

Transition from alive with coronary heart disease to death from coronary heart disease

Risk factor	Coefficient	Standard error	p-value
Systolic blood presure	-0.0099	0.0117	0.396
Body mass	0.0638	0.0750	0.395
Cholesterol	0.0072	0.0045	0.108
Smoking	-0.3858	0.4766	0.418

END

FILMED

12-84

DTIC